MRC Epidemiology Unit

Date created 7/3/2006

Some Problems in Genetics, Statistics and Computing

I will present some genetic problems and analytic tasks we received within the Unit. They are very interesting and have difficulties to various degrees. I will discuss their solutions and point out some references. I will also return to the access of databases as briefed at genetic meetings on Oct 27 and Dec 8 last year. This will be relevant to the genome-wide association studies. Finally, I will give a summary of the implications.

1. Some genetic problems

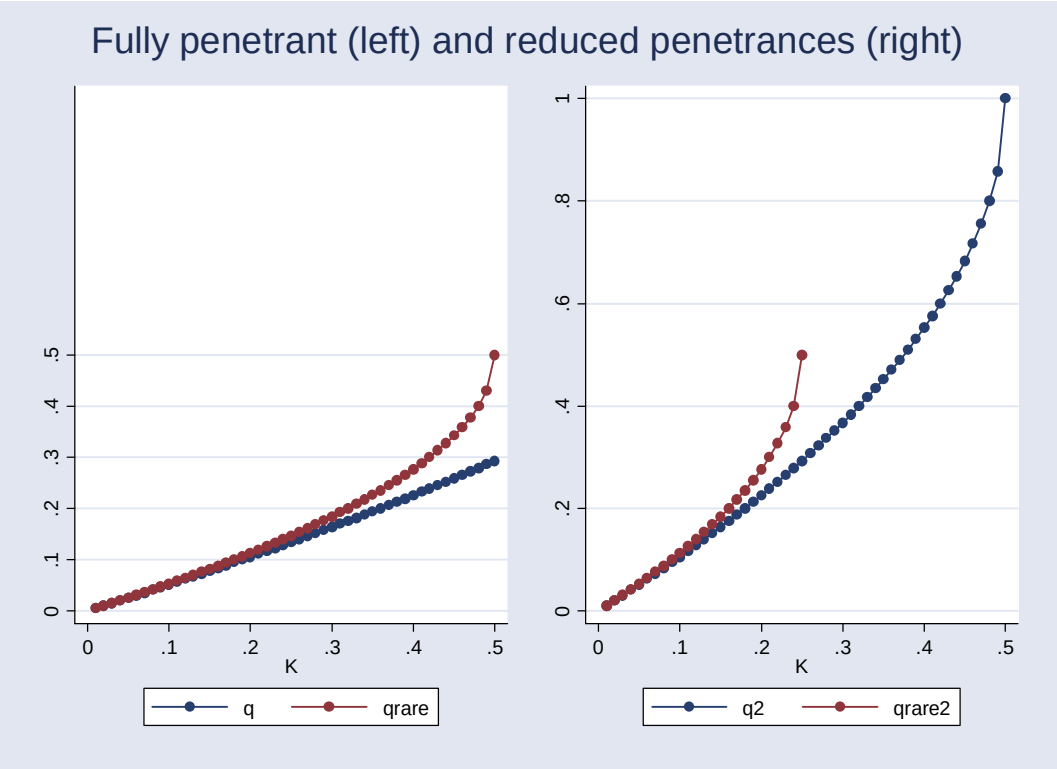
1) **Carrier frequencies** (from Liz)

The concept of disease model or mode of inheritance (MOI) is very important. Examples are "model-free" linkage analysis as implemented in computer programs MFLINK and GENEHUNTER/MODSCORE, where the focus is on the penetrance rather than allele frequencies considered here. We also used similar idea in combined linkage and linkage disequilibrium analysis of family data, allelic association of population data as implemented in the computer program EHPLUS. Other examples are power calculation for gene-environment interaction using computer program QUANTO.

Assuming the disease-predisposing locus is binary with alleles 1 (wild type) and 2 (mutant type), respectively. Let p = frequency of allele 1, q = frequency of allele 2, so $p + q = 1$. Let $f_i, i = 0,1,2$ be the associate penetrances, the population prevalence is then $K = f_0 p^2 + 2pqf_1 + q^2 f_2$. Under a dominant model, $f_0 = 0$, $f_1 = f_2 = f$, K is also the carrier frequency, the corresponding (mutant) allele frequency is $1-\sqrt{1-K/f}$, or $1-\sqrt{1-K}$ if the binary trait is also Mendelian. When the disease is very rare, K is dominated by the term for heterozygote, we then have $0.5-\sqrt{0.25-0.5K/f}$ instead. Under recessive model $f_0 = f_1 = 0$, $f_2 = 1$, this will be $\sqrt{K/f}$. We can examine the effect of rare versus common diseases with reduced or incomplete penetrance, e.g. $f < 1$ (non-Mendelian).

We now use Stata 8 for numerical comparison.

```
clear
set obs 50
gen K=_n*0.01
gen q=1-sqrt(1-K)
gen qrare=0.5-sqrt(0.25-0.5*K)
graph twoway (scatter q K, c(l) yscale(r(0 1)))  (scatter qrare K, c(l) yscale(r(0 1))), saving(full, replace)
gen f=0.5
gen K2=K/f
gen q2=1-sqrt(1-K2)
gen qrare2=0.5-sqrt(0.25-0.5*K2)
graph twoway (scatter q2 K, c(l)) (scatter qrare2 K, c(l)), saving(half, replace)
graph combine full.gph half.gph, title("Fully penetrant (left) and reduced penetrances (right)")
```



The right panel has $f = 0.5$ but only produces legitimate solutions up to 0.5.

2) **Power of allelic association** (from Manj)

Two power estimates have been obtained using GLM (QUANTO for InterAct) or semi-parametric model (case-cohort design for obesity genome-wide association), but here we show how to proceed for problems if some empirical data are available. The diagram is drawn to show a well-separated null and alternative. We also examine the power as a function of type I error. Again we use R, although this could easily be translated into MicroSoft Excel as originally intended.

```
par(mfrow=c(2,2))
df <- 10
ncp <- 15
curve(dchisq(x,df,0),0,50, ylab="Chi-squared with df=10")
curve(dchisq(x,df,ncp),0,50,add=T,col="blue")
mtext("Chi-squared distribution with df=10, ncp=0,15")

alpha <-  seq(0.0001,0.1,0.0001)
q <-  qchisq(1-alpha,df,0)
power <-  1 - pchisq(q,df,ncp)
plot(alpha,power,type="l")
mtext("The power as a function of type I error")
```

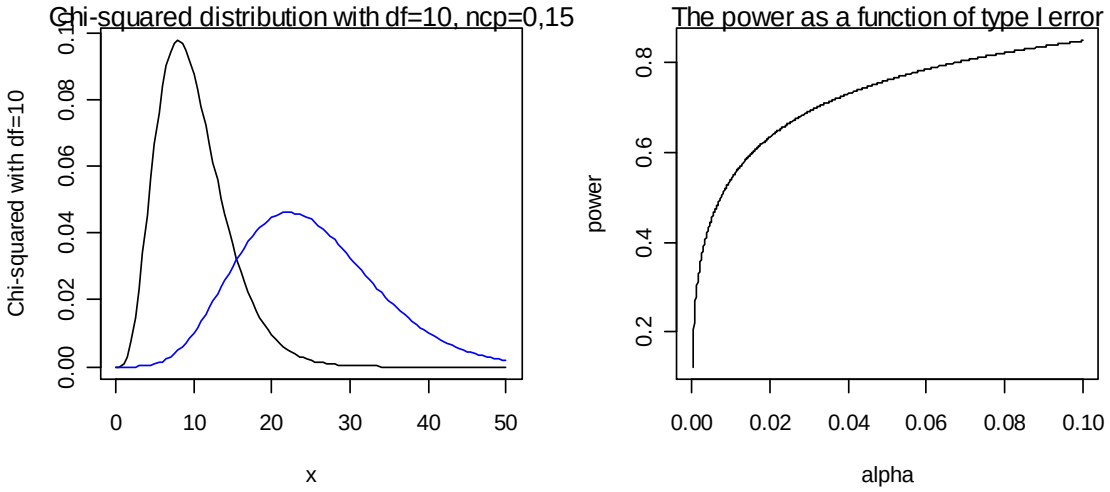
As in the case of χ^2 test in single-point analysis, one may have some idea of the effect size via the observed statistic. Given this is true, one can translate to estimate power or required sample size if the type I error is also given. For case-control data with a polymorphism with N alleles, df=N-1. We assume that a Chi-squared statistic of 2 is obtained on a tri-allelic system based on a sample of

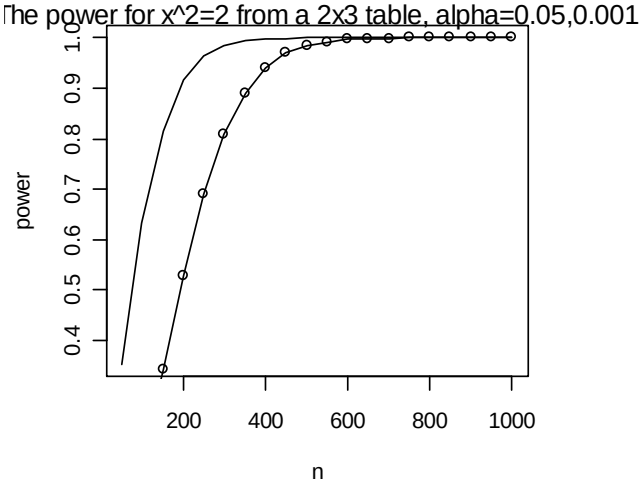
30 cases and controls. We wish to obtain power estimates for a range of sample sizes with significance levels 0.05 and 0.001.

```
x2obs <-  2
dfobs <-  2
nobs <-  30
n <-  seq(50,1000,50)
w <-  x2obs / nobs

alpha <-  0.05
q <-  qchisq(1-alpha,dfobs,0)
pexp <-  1 - pchisq(q, dfobs, w*n)
plot(n, pexp, type="l", ylab="power")

alpha <-  0.001
q <-  qchisq(1-alpha,dfobs,0)
pexp <- 1 - pchisq(q, dfobs, w*n)
lines(n, pexp, type="o")
mtext("The power for x^2=2 from a 2x3 table, alpha=0.05,0.001")
```



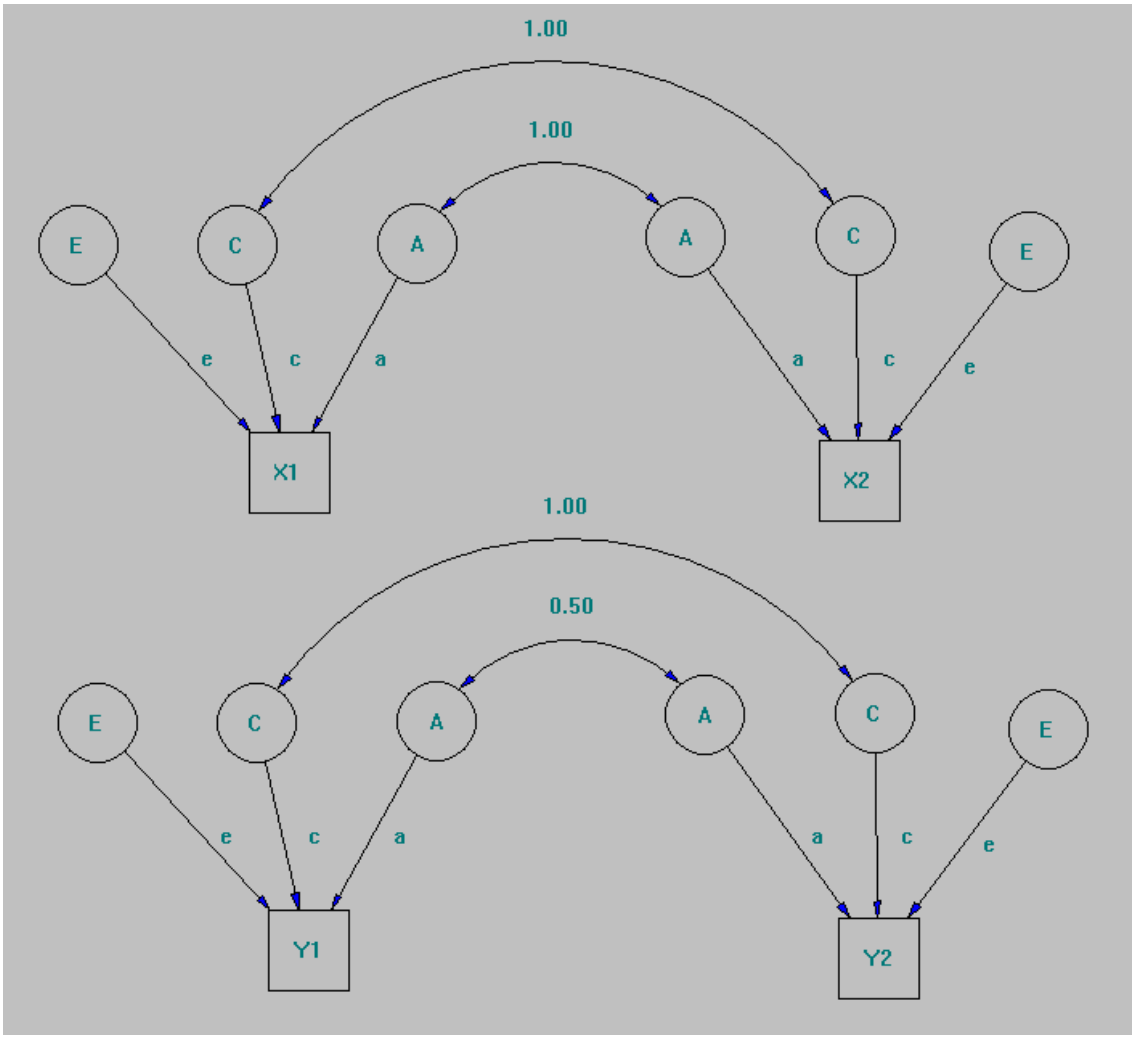


3) **Heritability estimate based on twins** (from Ken)

Twin studies are one of the major sources of heritability estimates and widely used in quantitative genetics. Workshops are held each year on twin data analysis. Briefly, what we observed are phenotypes for both MZ and DZ twin pairs, from which phenotypic correlations (r_{MZ}, r_{DZ}) can be obtained. We consider the ACE model as an example shown via a path diagram below.

One can easily derive the covariance or correlation between two types of twin pairs, but an intuitive way is to follow the path-tracing rules. Both approaches give the same answer, and for MZ twin pairs, $r_{MZ} = a^2 + c^2$, and for DZ twin pairs, $r_{DZ} = 0.5a^2 + c^2$ and therefore $a^2 = 2(r_{MZ} - r_{DZ})$. Assuming the phenotypes are standardised, this is also the usual heritability estimate. As we can see it requires the assumption of common environment between MZ and DZ twins. Alternatives are available (Feldman et al. 1997, Science) and more assumptions are required (Falconer and Mackay 1996). Its variance of can be obtained, based on that of correlation, i.e., $4\left[\left(1-r_{MZ}^2\right)^2/n_{MZ} + \left(1-r_{DZ}^2\right)^2/n_{DZ}\right]$. The model can be extended to consider gene-environmental interactions (Guo 2000, Hum Hered).

It is customary to consider a range of models including ADE, AE, E, etc, using programs such as TWINAN90 (Williams et al. 1992, CMPB), LISREL (Neale and Cardon 1992), and Mx (Neale and Maes 2004). A recent review of twin studies is available (Boomsma et al. 2002, Nat Rev Genet).



The ACE model of twins (top=MZ twin, bottom=DZ twin)

4) **Haplotype trend regression** (from Paul)

This method builds the counts for individual haplotypes to be used in a prospective model. Its versatile nature allows for a wide range of statistical models such as GLM and GLMM or even GLLAMM.

SAS/GENETICS provides a sketch how the count matrix could be built but lacks of a generalisation. It thus avoids the need for programs such HTR (Zaykin et al. 2002, Hum Hered) and tagSNPs (Stram et al. 2003, Hum Hered).

The SAS macro is available from V:\ drive. A version using GENECOUNTING is available as a function in R.

5) **EPIC 400 data** (from EPIC)

Analyses were carried out in SAS and graphics in S-PLUS, all under Linux. The analyses include HWE → single point analysis → **Haplotype analysis**, and adjustment for multiple testing including FDR and permutation tests. Other analysis would involve examination of population structure, linkage disequilibrium and recombination analysis, as well as comparison of findings with the literature, e.g. region in NCBI.

This is scheduled to be reported at the work in progress meeting early May.

6) **The relative performance of LD** (from Bert)

Although there is a large literature it remains a problem!

7) **An optimisation problem** (from Soren)

This is a computer-intensive task. An attempt has been made through computer simulation in C followed by a reformulation as a optimisation problem implemented in SAS/IML. Both give stable results. It poses the problem of seeking global optimum, a generic problem and has big potential in other applications.

8) **Estimation of age-adjusted relative risk** (from Nita)

The following is the actual exchange of e-mails. The expressions were illustrated with OR (as it is better known and an approximation for RR).

Q

Is it "correct" to apply a Mantel-Haenszel adjustment to a relative risk calculated in a prospective study (I have only used it before for odds ratios adjusted for age from case control study data)? We have some data from a table in a paper which presents RR for the association between baseline diabetes (yes vs no) and incident pulmonary tuberculosis by age group, and from this table I want to calculate the age-adjusted RR.

So I have for example –

Age RR smear positive TB

30-39 6.63 (1.65-26.63)
40-49 12.68 (7.44 – 21.61)
50-59 5.18 (3.07 – 8.73)
60+ 3.96 (1.38 – 11.36)

I cannot easily get the age stratum specific 2x2 tables to get the "a b c d" for working out the weights for each stratum, but the 95% CIs are available as above. The paper presents a crude RR estimate for whole population of 6.92 (4.96 – 9.65), but I wanted to be able to get an age adjusted one.

Can you think of a way of doing so with the data above?

A

It seems the first part is easy; the answer is YES. With the second part, I should try to formula some nonlinear simultaneous equations – as you can see we have four tables each containing a_i, b_i, c_i, d_i as unknowns, we can follow classic formulae such as ad/bc, 1/a+1/b+1/c+1/d and their variations, plus those for A, B, C, and D. Essentially, after transformations we have RR_i, Var(RR_i) for each table and RR and Var(RR) for the combined. The system is undetermined since we have some extra unknowns. This is an ideal problem for symbolic computation for Mathematica, Maple or Matlab that I am not quite experienced to program them – or we have some iterative numerical procedure, e.g. we can run some grids for the unknown. Finally, with these estimates we can run a logistic regression such as logit(y)=a+b TB + c age?

I hope we can get to the bottom of the problem some time soon.

2. Large data experiment

The following is a short report of some experiments I have conducted in preparation of analysis of large genomic data. This is useful since we will work on genetic database containing hundreds of thousands of single-nucleotide polymorphisms (SNPs) along with other biological and environmental covariates. The experiments involve MySQL (http://www.mysql.com) database accessible from S-PLUS and R on our Linux Box (131.111.251.106), but ODBC (open database connectivity) works with R, SAS, S-PLUS and Stata. The programs for C/Perl/PHP/Python are relevant if we are going to provide some interface for internal or public access of our data.

Although the examples are for MySQL, other databases such as Oracle should work similarly. The ODBC mechanism is quite general in the sense that it is not only applicable to Linux documented here but also to Windows system. Adding R/S packages under Windows have been described earlier (V:\ drive), but I did not try MySQL with S-PLUS under Windows. Nevertheless, RODBC should work with R. Furthermore, it is possible to communicate between Windows and Linux database server.

Setup

A file called .odbc.ini and containing the following lines is required at the $HOME directory.

```
[ODBC Data Sources]
myodbc      = MyODBC 2.50 Driver DSN

[myodbc]
Driver       = /usr/lib/libmyodbc.so
Description  = MyODBC 2.50 Driver DSN
SERVER       = localhost
PORT         =
```

```
USER         = username
Password     =
Database     = test
OPTION       = 3
SOCKET       =

[Default]
Driver       = /usr/lib/libmyodbc.so
Description  = MyODBC 2.50 Driver DSN
SERVER       = localhost
PORT         =
USER         =  username
Password     =
Database     = test
OPTION       = 3
SOCKET       =
```

Several drivers can be listed but with default section is used.

A sample session

We log on to a MySQL session with the command,

%mysql -u username -p

We first use the following SQL script to generate a database called test containing a table called pet.

```
CREATE DATABASE test;
USE test;
CREATE TABLE pet (name VARCHAR(20), owner VARCHAR(20),species VARCHAR(20), sex
CHAR(1), birth DATE, death DATE);
SHOW TABLES;
DESCRIBE pet;
INSERT INTO pet VALUES ('Puffball','Diane','hamster','f','1999-03-30',NULL);
DESCRIBE pet;
LOAD DATA LOCAL INFILE 'pet.txt' INTO TABLE pet LINES TERMINATED BY '\r\n';
INSERT INTO pet VALUES ('Puffball','Diane','hamster','f','1999-03-30',NULL);
SELECT * FROM pet;
```

e.g., by the source command within MySQL. The following file, named pet.txt, at the current directory can be imported into the table.

```
Fluffy   Harold   cat     f     1993-02-04
Claws    Gwen     cat     m     1994-03-17
Buffy    Harold   dog     f     1989-05-13
Fang     Benny    dog     m     1990-08-27
Bowser   Diane    dog     m     1979-08-31     1995-07-29
Chirpy   Gwen     bird    f     1998-09-11
Whistler Gwen     bird          1997-12-09
Slim     Benny    snake   m     1996-04-29
```

% mysqlimport --local test pet.txt

SAS

The MySQL engine requires SAS/ACCESS to be installed.

```
libname test odbc datasrc=myodbc user=username;
proc print data=test.pet;
run;
proc sql;
    connect to odbc as test
    (datasrc=myodbc user=uername);
    select * from test.pet;
libname test2 mysql database=test user=username;
proc print data=test2.pet;
run;
proc sql;
    connect to mysql as test2
    (database=test user=username);
    select * from test2.pet;
```

It is necessary to allocate enough space for SAS system working directory (e.g. /tmp), or not using it at all by specifying

```
SAS  -work $HOME/work
```

in order to allocate sufficient sorting space.

Stata

Commands for accessing ODBC through Stata is given as follows,

```
odbc list
odbc query "myodbc"
odbc load, exec("select * from pet")
```

R

Within R, both RMySQL and RODBC can be set up to access MySQL databases. Example code is given as follows for RMySQL,

```
library(RMySQL)
m <- dbDriver("MySQL")
con <- dbConnect(m,"test")
rs <- dbSendQuery(con,"select * from pet")
df <- fetch(rs,n=3)
df
```

The appropriate R functions can also write data directly into MySQL.

S-PLUS

The S-MySQL package is by David James and available from http://stat.bell-labs.com/RS-DBI/index.html.

```
library("S-MySQL",lib.loc="/$HOME/S/library")

# initialize S/Splus as a MySQL client
  mgr <- dbManager("MySQL")
```

```
# create a connection to a MySQL server
  con <- dbConnect(mgr, user="username", dbname="test")
# run a query, leave results on the server
  rs <- dbExec(con, "select * from pet")
# fetch up to, say, 50 records
  df <- fetch(rs, n = 5)
  # close resultSet rs and connection con
  close(rs)
  close(con)
```

S-PLUS 7 under Linux does not support ODBC but it does under Windows.

C

```
*cc -I/usr/include/mysql -L/usr/lib/mysql -lmysqlclient -lnsl -lm -lz -o pet
pet.c*/

#include <mysql.h>
#include <stdio.h>

main() {
   MYSQL *conn;
   MYSQL_RES *res;
   MYSQL_ROW row;

   char *server = "131.111.251.106";
   char *user = "username";
   char *password = "password";
   char *database = "test";

   conn = mysql_init(NULL);

   /* Connect to database */
   if (!mysql_real_connect(conn, server,
         user, password, database, 0, NULL, 0)) {
      fprintf(stderr, "%s\n", mysql_error(conn));
      exit(0);
   }
   /* send SQL query */
   if (mysql_query(conn, "SELECT * FROM pet")) {
      fprintf(stderr, "%s\n", mysql_error(conn));
      exit(0);
   }
   res = mysql_use_result(conn);
   /* output fields 1 and 2 of each row */
   while ((row = mysql_fetch_row(res)) != NULL)
      printf("%s %s\n", row[1], row[2]);
   mysql_close(conn);
}
```

Under Solaris, it is necessary to specify -lsocket as well.

There are other possibilities, such as Perl, PHP and Python.

3. Summary

The most striking thing to me is that some seemingly trivial problems may not always be straightforward, and quite often require serious commitment. These are the very motivation for further work.

The list of problems is not exhaustive and ready to be extended. The large data experiments already revealed some limitations of our current computing environment; the remaining work would involve Linux cluster, the system for reliable, multitasking and parallel computing.

These problems depict a broad picture of research areas. The statistical tools available to us are excellent, and many pieces are in place for data access and analysis with just slight more effort. Two implications seem clear:

- It is useful to have a repository of tools and methods for data analysis (e.g. a workbook?) and preferably involve most people.

- It will be necessary to organise some training session for specific problems or projects, e.g. the use of Linux system and application of statistical packages. Some basics of Linux system, including customised procedures, will be needed.

As with Linux, the following routines have been established so far,

- puTTY access to the Linux Box including X-Windows services

- WinSCP for remote copy

- SAS 9.1, S-PLUS 7, Stata 9.1, STAT/Transfer 8, R 2.2.1

Virtual network computing (VNC) allows for remote connection from office or home. Other options include cygwin (http://cygwin.com), which might also serve as an intermediate step between Windows and Linux systems.

There is a barrier is to be overcome – most users are currently tied to Windows. Although one may be unwilling to invest time on some aspects, collectively they may turn to be very helpful in general. With participation of many individuals, it would be mutually beneficial and easier than expected.